The inaccuracy of patient recall for COPD exacerbation rate estimation and its implications: results from central adjudication

Frei, Anja ; Siebeling, Lara ; Wolters, Callista ; Held, Leonhard ; Muggensturm, Patrick ; Strassmann, Alexandra ; Zoller, Marco ; Ter Riet, Gerben ; Puhan, Milo A

Abstract: BACKGROUND: COPD exacerbation incidence rates are often ascertained retrospectively through patient recall and self-reports. We compared exacerbation ascertainment through patient self-reports and single-physician chart review to central adjudication by a committee and explored determinants and consequences of misclassification. METHODS: Self-reported exacerbations (event-based definition) in 409 primary care patients with COPD participating in the International Collaborative Effort on Chronic Obstructive Lung Disease: Exacerbation Risk Index Cohorts (ICE COLD ERIC) cohort were ascertained every 6 months over 3 years. Exacerbations were adjudicated by single experienced physicians and an adjudication committee who had information from patient charts. We assessed the accuracy (sensitivities and specificities) of self-reports and single-physician chart review against a central adjudication committee (AC) (reference standard). We used multinomial logistic regression and bootstrap stability analyses to explore determinants of misclassifications. RESULTS: The AC identified 648 exacerbations, corresponding to an incidence rate of 0.60 ± 0.83 exacerbations/patient-year and a cumulative incidence proportion of 58.9%. Patients self-reported 841 exacerbations (incidence rate, 0.75 ± 1.01; incidence proportion, 59.7%). The sensitivity and specificity of self-reports were 84% and 76%, respectively, those of single-physician chart review were between 89% and 96% and 87% and 99%, respectively. The multinomial regression model and bootstrap selection showed that having experienced more exacerbations was the only factor consistently associated with underreporting and overreporting of exacerbations (underreporters: relative risk ratio [RRR], 2.16; 95% CI, 1.76-2.65 and overreporters: RRR, 1.67; 95% CI, 1.39-2.00). CONCLUSIONS: Patient 6-month recall of exacerbation events are inaccurate. This may lead to inaccurate estimates of incidence measures and underestimation of treatment effects. The use of multiple data sources combined with event adjudication could substantially reduce sample size requirements and possibly cost of studies. CLINICAL TRIAL REGISTRATION: www.ClinicalTrials.gov, NCT00706602.

DOI: https://doi.org/10.1016/j.chest.2016.06.031

Originally published at:
Frei, Anja; Siebeling, Lara; Wolters, Callista; Held, Leonhard; Muggensturm, Patrick; Strassmann, Alexandra; Zoller, Marco; Ter Riet, Gerben; Puhan, Milo A (2016). The inaccuracy of patient recall for COPD exacerbation rate estimation and its implications: results from central adjudication. Chest, 150(4):860-868.
DOI: https://doi.org/10.1016/j.chest.2016.06.031

Word counts
Abstract: 247
Text: 2562

The Inaccuracy of Patient Recall for COPD Exacerbation Rate Estimation and its Implications: Results from Central Adjudication

Results from Central Adjudication of Exacerbations

Anja Frei, PhD[1]*; Lara Siebeling, PhD[2]; Callista Wolters, MD[1]; Leonhard Held, Prof[1]; Patrick Muggensturm, MD[1]; Alexandra Strassmann, MSc[1]; Marco Zoller, MD[3]; Gerben ter Riet, PhD[2+] & Milo A. Puhan, Prof[1+]

[+] Contributed equally as senior authors

[1]Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Switzerland
[2]Department of General Practice, Academic Medical Center, University of Amsterdam, the Netherlands
[3]Institute of General Practice and Health Services Research, University of Zurich, Switzerland

* Anja Frei, Epidemiology, Biostatistics and Prevention Institute
University of Zurich, Hirschengraben 84, 8001 Zürich, Switzerland
anja.frei@uzh.ch

Funding: This study is part of the prospective ICE COLD ERIC cohort study which was supported by the Swiss National Science Foundation (grant # 3233B0/115216/1), Lung Foundation Netherlands (grant # 3.4.07.045), Stichting Astmabestrijding (grant # SAB 2012/043 and Zurich Lung League (unrestricted grant).

Conflict of interest: The authors declare that they have no conflicts of interests.

Prior abstract publication: Parts of the results from this manuscript have been presented at the European Respiratory Society Annual Congress 2012 in Vienna.

Abbreviations list

AC = Adjudication committee

ATS = American Thoracic Society

CI = Confidence interval

COPD = Chronic obstructive pulmonary disease

CRQ = Chronic Respiratory Questionnaire

ERS = European Respiratory Society

FEV_1 = Forced expiratory volume in 1 second

FT = Feeling Thermometer

FVC = Forced vital capacity

GOLD = Global Initiative for Chronic Obstructive Lung Disease

HADS = Hospital Anxiety And Depression Scale

ICE COLD ERIC = International collaborative effort on chronic obstructive lung disease: exacerbation risk index cohorts

LABA = Long-acting beta-agonists

OR = Odds ratio

RCT = Randomised controlled trial

RRR = Relative risk ratio

TORCH = Towards a Revolution in COPD Health

UPLIFT = Understanding Potential Long-term Impacts on Function with Tiotropium

ABSTRACT

Background

COPD exacerbation incidence rates are often ascertained retrospectively, through patient recall and self-reports. We compared exacerbation ascertainment through patient self-reports and single physician chart review to central adjudication by a committee and explored determinants and consequences of misclassification.

Methods

Self-reported exacerbations (event-based definition) in 409 primary care COPD patients participating in the ICE COLD ERIC cohort were ascertained 6-monthly over 3 years. Exacerbations were adjudicated by single experienced physicians and an adjudication committee who had information from patient charts. We assessed the accuracy (sensitivities and specificities) of self-reports and single physician chart review against a central adjudication committee (reference standard). We used multinomial logistic regression and bootstrap stability analyses to explore determinants of misclassifications.

Results

The adjudication committee identified 648 exacerbations, corresponding with an incidence rate of 0.60±0.83 exacerbations/patient-year and a cumulative incidence proportion of 58.9%. Patients self-reported 841 exacerbations (incidence rate 0.75±1·01, incidence proportion 59.7%). Sensitivity/specificity of self-reports were 84%/76%, those of single physician chart review between 89-96% and 87-99%. The multinomial regression model and bootstrap selection showed that having experienced more exacerbations was the only factor consistently associated with under- and over-reporting of exacerbations (under-reporters: relative risk ratio 2.16, 95% CI 1.76-2.65; over-reporters: relative risk ratio 1.67, 95% CI 1.39-2.00).

Conclusions

Patient 6-month recall of exacerbation events are inaccurate. This may lead to inaccurate estimates of incidence measures and underestimation of treatment effects. The use of multiple data sources combined with event adjudication could substantially reduce sample size requirements and possibly cost of studies.

Clinical Trial Registration: www.ClinicalTrials.gov, NCT00706602

INTRODUCTION

Good measurement is central to good science. Similarly, accurate ascertainment of endpoints in observational studies and randomised controlled trials (RCTs) is crucial to minimise endpoint misclassification. Concerns about misclassification of outcomes prompted the use of endpoint adjudication committees (ACs), particularly in large cardiovascular RCTs.[1,2] The value of endpoint ACs has been debated recently because they require considerable investments and some studies found no discrepancy between effect estimates based on centrally adjudicated endpoints or adjudications made by single local investigators.[3–5] However, in studies assessing misclassification of different types of outcomes, the degree of misclassification increased with increasing ambiguity (subjectiveness) of the outcome at issue.[6]

There is little evidence on when to use outcome ACs in studies with chronic obstructive pulmonary disease (COPD) patients. Candidate outcomes likely to benefit from adjudication by a central committee in COPD studies include cause-specific mortality and exacerbations. The Towards a Revolution in COPD Health (TORCH)[7] and Understanding Potential Long-term Impacts on Function with Tiotropium (UPLIFT)[8] trials had ACs for cause-specific mortality and found substantial disagreement between local investigators and ACs.[9,10] COPD exacerbations have rarely been adjudicated centrally by experts blinded to treatment or exposures,[11,12] despite their importance as an outcome measure.[13–16] Ascertainment of exacerbations is challenging because of mimicking differential diagnoses (e.g. worsening of heart failure, pulmonary embolism), since the cause of exacerbations often cannot be determined and because several sources of information (patient self-reports, patient charts, emergency healthcare visits) are needed to avoid missing or misclassifying exacerbations.

Exacerbation rates for enrolment or evaluation in clinical trials are usually ascertained through patient recall or single physicians who review the available information about a

patient and make judgments on the occurrence of exacerbations (expert adjudication). However, little is known about the accuracy of these reports. Misclassification of exacerbations may lead to biased effect estimates. The exact nature of the bias may be different for (risk) ratio measures and difference measures and can also depend on whether exposure is misclassified (noncompliance in trials).[17,18] Our aim was to evaluate the accuracy of the ascertainment of COPD exacerbations through patient recall and self-reports and single physician chart reviews against a reference standard of a central AC. In addition, we explored predictors of inaccurate ascertainment of COPD exacerbations. Finally, we explored potential consequences of misclassification for treatment effects in trials and sample size requirements.

MATERIAL AND METHODS

The unabridged version of the Material and Methods section is presented in e-Appendix 1.

Study design and study subjects

This study was nested within the prospective International collaborative effort on chronic obstructive lung disease: exacerbation risk index cohorts (ICE COLD ERIC) and comprised 3 years. 409 primary care patients (≥40 years of age) from the Netherlands (n=258) and Switzerland (n=151) with COPD (post-bronchodilator forced expiratory volume in 1 second [FEV_1]/forced vital capacity [FVC]<0.7, FEV_1<80% predicted) were included. All patients provided written informed consent. The study has been approved by the ethics committees in Zurich (EK-1519) and Amsterdam (MEC-08-073). Detailed information on the study design[19] and results[20–22] were published elsewhere.

Definition and ascertainment of exacerbations

We used an event-based definition of exacerbations that required 1) an unscheduled physician contact in a hospital, private practice or by telephone for worsening of dyspnoea, cough, increased sputum production and/or a change in sputum colour AND 2) an electronic or hand-written documentation in the patient record of a new prescription or dosage increase of systemic steroids and/or new prescription of an antibiotic.[19,23]

We used three different methods to ascertain exacerbations. First, through patient recall and self-reports, with data gathered every six months through personal interview by experienced study nurses. Second, through independent review of patient charts and study case report forms by seven single experienced physicians. These experts individually and independently decided about occurrence of exacerbations, onset date (prescription/dosage increase or first pill taken) and treatment setting. To distinguish between new exacerbations and slow-to resolve ones or relapses, an interval of at least one month was required. Third,

through central AC meetings where a group of experts (formed by the independent reviewers) reached final consensus on exacerbations for each patient (reference standard). The AC meetings were organised by study staff that neither participated in the discussions nor had influence on decisions made. The expert committee discussed instances where decisions on exacerbations were discrepant, based on their individual decisions, re-review of patient charts and case report forms.

Analyses

We tabulated total number of recalled and self-reported exacerbations against total number of centrally adjudicated exacerbations at the level of patients. We calculated exacerbation incidence rates and cumulative incidence proportions as self-reported by patients, as decided by individual experts (for incidence proportions) and as adjudicated in the AC. We calculated sensitivity and specificity of patient self-reports and independent review of patient charts and study case report forms by single physicians compared to the reference standard of centrally adjudicated exacerbations (categorised: no exacerbation/≥1 exacerbations). To explore predictors of misclassification, we generated a patient-specific rate difference reflecting the difference between self-reported and adjudicated exacerbation rate and categorised the variable into correct self-reports, over- and under-reports (predictors: sex, age, education, living situation, working status, COPD severity, health status (feeling thermometer)[24,25], quality of life (Chronic Respiratory Questionnaire),[26] depression/anxiety (Hospital Anxiety And Depression Scale),[27] number of comorbidities, previous exacerbations). We used multinomial logistic regression and bootstrap stability analyses. We conducted all analyses using Stata Statistical Software version 13.1, StataCorp LP, College Station, TX.

RESULTS

Patient characteristics and adjudicated exacerbations

57.5% (233/409) of the patients were male. Overall mean age was 67.3±10.0 years, FEV_1 in % predicted 55.4±16.6 and MRC score 1.9±1.5. 261 (63.8%) patients were classified in Global Initiative for Chronic Obstructive Lung Disease (GOLD) stage II, 89 (21.8%) in GOLD stage III and 59 (14.4%) in GOLD stage IV.

In total, across all 409 patients, the AC identified 648 exacerbations during the three study years. 94% (n=606) of the exacerbations were outpatient-treated, 6% (n=38) were treated in a hospital setting and for four exacerbations the treatment setting was unknown. 168 (41.1%) patients did not have an exacerbation during the study time, 101 patients (24.7%) had one, 41 patients (10.0%) had two, 34 (8.3%) patients had three and 65 patients (15.9%) had four or more exacerbations. The incidence rate was 0.60±0.83 exacerbations/patient-year and the incidence proportion (having experienced at least one exacerbation) 58.9% (n=241). The total person-time at risk was 1153.4 years and the mean length of follow-up was 2.82±0.49 years.

Accuracy of exacerbations

The patients recalled and self-reported 841 exacerbations over three years. Incidence rate of self-reported exacerbations was 0.75±1.01 exacerbations/patient-year and incidence proportion 59.7% (n=244). The total number of exacerbations during the three study years was correctly reported by 48% (196/409) of the patients; 65% of the patients who correctly reported the number of exacerbations had no exacerbations (127/196). The proportion of over-reports on exacerbations was higher than the proportion of under-reports (139/409, 34% vs. 74/409, 18%). Thus there was more false positive than false negative misclassification (Table 1).

Using the centrally adjudicated exacerbations as a reference standard, patient recall and self-reports of exacerbations achieved a sensitivity of 84% and a specificity of 76%. Independent review of patient charts and study case report forms by the 7 single physicians achieved sensitivities ranging from 89-96% and specificities ranging from 87-99% (Table 2).

Determinants of misclassification

Table 3 shows the results of the multinomial regression model. "Number of previous exacerbations" (adjudicated) and "feeling thermometer" were the only predictors for which the inclusion fraction exceeded our 67% threshold; for number of previous exacerbations in 100%, for feeling thermometer in 67.7%. Compared to patients who correctly reported the total number of exacerbations, patients who under- or over-reported experienced a higher number of exacerbations (under-reporters: relative risk ratio [RRR] 2.16, 95% confidence interval [CI] 1.76-2.65; over-reporters: RRR 1.67, 95% CI 1.39-2.00) (RRR describes the multiplicative effect of a unit increase in each predictor on the odds of over- or under-reporting instead of correctly self-reporting exacerbations [base category]). The feeling thermometer was not associated with over- or underreporting (under-reporters: RRR 0.99, 95% CI 0.97-1.02; over-reporters: RRR 0.99, 95% CI 0.97-1.01). The reduced multinomial regression model that includes only these two selected variables is presented in e-Table 1.

DISCUSSION

Our study showed that patient 6-month recall of COPD exacerbations, the most commonly used method to assess exacerbation rates in COPD research, was fairly inaccurate when compared against an AC as the reference standard. Having experienced a higher number of exacerbations in the past was the only predictor associated with misclassification. The independent assessments of comprehensive patient information on exacerbations from patient charts and study case report forms by single experienced physicians were more accurate than patient self-reports only and, for some physicians, approximated the accuracy of adjudication by an AC.

In general, patients tended to over-report more than to under-report exacerbations. The AC's feedback showed that a major reason for over-reporting was that patients correctly remembered physician encounters or hospitals visits, but that many of these events were unrelated to COPD exacerbations. For example, patients received antibiotics for urinary tract infection or were admitted to hospitals because of an injury. These events were clearly documented in the patient charts of the general practitioners. This highlights the advantage of going beyond patient self-reports for outcomes that are difficult to ascertain, such as COPD exacerbations. Some misclassifications concerned differential diagnoses such as pulmonary embolism, and some over-reporting occurred because patients referred to new exacerbations whereas the AC did not classify them as new but as slow recoveries from previous exacerbations.

Little is known so far about the accuracy of patient recall of exacerbation events and single physician judgements based on patient chart reviews, the two most commonly used methods to ascertain COPD exacerbation rates. A recently published study used medical records and two blinded investigators to verify patient reports on exacerbations but did not report on agreement of expert adjudication with patient self-reports.[12] Investigators of the Women's Health Initiative trial found that sensitivity of patient recall and self-reports on

cardiovascular events, compared to single expert assessment, was highly dependent on the type of outcome. Self-reported events like angina, peripheral vascular disease and congestive heart failure attained substantially lower sensitivities (38-49%) when compared against central adjudication than "hard" endpoints such as coronary bypass surgery or angioplasty (84-90%). The accuracy of single expert judgements compared to central adjudication was higher, with sensitivities/specificities similar to those observed in our study.[6] These findings support the assumption that the value of ACs increases for outcomes that are experienced by patients as less determinate compared to more distinct events such as surgeries. Patients probably experience outpatient-treated exacerbations as changes of their underlying disease symptomatology and pharmacologic therapy. Therefore, exacerbation events offer less anchors in time and patients will presumably have more difficulties recalling them in 6-months interviews. Assessed retrospectively through patient recall and self-reports, they are more prone to misclassification, particularly in patients with frequent exacerbations, multiple co-morbidities and older age.

One crucial consequence of endpoint misclassifications is their impact on treatment effect estimates and, consequently, sample size requirements. Although non-differential misclassification (of a binary outcome) often leads to underestimation of true treatment effects, this is not always the case.[18,28] While these phenomena are often well appreciated in observational research, there is still sparse awareness in RCTs that inaccurate measurement may have a serious impact on validity and precision of effect estimates.[17]

To illustrate the potential impact of treatment effect estimates misclassification and sample size requirements for RCTs using the outcome exacerbation incidence proportion, we repeated a meta-analysis of RCTs on long-acting beta-agonists (LABAs) as first-line maintenance therapy vs. placebo.[29] Like in our study, exacerbations were assessed using patient self-reports and event-based definition. We assumed several exemplary scenarios for misclassifications, using sensitivities/specificities of a) 80%/70%, b) 84%/76% (as our study

detected), c) 80%/85%, d) 95%/85%. We recalculated the meta-analysis adjusting for the assumed examples using *logitem* command (Stata Statistical Software version 13.1). The pooled odds ratio of LABA versus placebo decreased (i.e. resulted in larger treatment effects) from 0.81 (95% CI 0.75-0.88) as reported in the original study to a) 0.59 (95% CI 0.43-0.81), b) 0.65 (95% CI 0.52-0.81), c) 0.69 (95% CI 0.58-0.81), d) 0.73 (95% CI 0.63-0.84). To delineate the consequences of AC deployment for sample size requirements in future trials, we calculated sample sizes using the original published and corrected ORs. Depending on baseline risk for exacerbations and degree of expected misclassifications, elimination of exacerbation misclassification led up to 2- to 6- fold reductions in required sample (Figure 1; e-Appendix 2).

The impact of misclassification on sample size requirements may be the same when using exacerbation incidence rates. In a simulation model derived from the misclassification based on ICECOLDERIC's ACs, we found that the incidence rate ratio was unbiased but the standard error of the incidence rate ratio was substantially inflated (i.e. less precision, e-Appendix 3).

In light of the available evidence and results of our study, we recommend the use of expert ACs in clinical trials and observational studies in which COPD patients are at high risk for exacerbations and in which COPD exacerbations are a key outcome. If it is deemed infeasible to centrally adjudicate exacerbations or in low-risk populations, we recommend single expert adjudication of information on exacerbations from case report forms, patient charts and other sources if available. If only patient self-reports on exacerbations are available and no additional patient charts assessment can be conducted, we recommend frequent (at least monthly) collection of patient recall and self-reports on exacerbations. Finally, the use of an AC, through its reduction of misclassification, leads to substantially larger or more precise effect estimates, thus greatly reducing sample size requirements for RCTs. This may lead to substantial pay-offs in terms of feasibility and cost of RCTs and observational studies.

Strengths of the study include that our population represents a large and diverse group of COPD patients from primary care. We used a clearly specified definition of exacerbations. Exacerbations were very carefully assessed by experienced and well-trained study nurses in patient friendly language. Also the ACs were carefully conducted in a standardised way by experienced physicians.

A limitation of our study is that even though we put great effort in accurate assessment and adjudication of exacerbations, we still may have missed or misclassified some. No standardised definition and measurement methods exist and different definitions may lead to different results.[30,31] We used an event-based definition, the most commonly used definition in drug development and clinical trials which is used by FDA and EMA to approve new therapies. This definition required documented worsening of symptoms and dosage increase/new prescription of systemic corticosteroids and/or antibiotics. We therefore accepted to miss mild exacerbations (ATS/ERS Task Force[23]) which involve an increase in respiratory symptoms that can be controlled by increase of usual medication. Our time criterion of ≥1 month between events is arbitrary to some extent. We used this interval because recent EXACT-PRO instrument data showed that it often takes patients weeks to recover from an exacerbation.[32] Finally, the sensitivities/specificities we provide for single physicians may be somewhat inflated because these experts were members of the AC, thereby violating the strict criterion that index test and reference standard have to be independent.

In conclusion, 6-month recall of exacerbation events are inaccurate when compared against an AC, leading to imprecise estimates of incidence rates and incidence proportions and underestimation of treatment effects. The use of several data sources combined with event adjudication could substantially reduce sample size requirements and possibly cost of studies.

Acknowledgements

Guarantor statement

AF had had full access to all the data in the study and had final responsibility for the manuscript, including the data and analysis.

Author contributions

AF, LS, CW, LH, PM, AS, MZ, GtR and MAP contributed to conception and design of the study. PM, LS and MZ contributed to patient recruitment, patient enrolment and data collection of the cohort study. MAP, GtR, LS, CW and AF designed, planned and piloted the AC process. LS, CW and GtR administratively guided through the AC meetings without participating in the discussion or influencing the decisions made. AF, CW and LS assessed and verified the data. LH developed the simulation model to re-estimate treatment effects based on extent of misclassification. AF, MAP, GtR, AS and LS contributed to statistical analysis. All authors contributed to interpretation of data. AF, MAP and GtR drafted the manuscript. All authors critically revised draft versions of manuscripts and approved the final version.

Declaration of interests

The authors declare that they have no competing interests.

Other contributions

We thank Ursula Schafroth (Horten Centre for patient-oriented research, University of Zurich, Switzerland) and Alice Karsten (Academic Medical Center, Department of General Practice, University of Amsterdam, the Netherlands) and the participating general practitioners and COPD patients in Switzerland and the Netherlands (Stichting Gezondheidscentra Amsterdam Zuidoost and Zorggroep Almere) who made this study possible by their enthusiastic participation. We also would like to thank all experts who

participated in the AC meetings in both countries. Finally, we thank Julia Braun-Grübel and Sarah Haile from the Epidemiology, Biostatistics and Prevention Institute, University of Zurich, for their statistical support.

References

1. Mahaffey KW, Harrington RA, Akkerhuis M, et al. Systematic adjudication of myocardial infarction end-points in an international clinical trial. *Curr Control Trials Cardiovasc Med* 2001;2(4):180–186.

2. Connolly S, Pogue J, Hart R, et al. Clopidogrel plus aspirin versus oral anticoagulation for atrial fibrillation in the Atrial fibrillation Clopidogrel Trial with Irbesartan for prevention of Vascular Events (ACTIVE W): a randomised controlled trial. *Lancet* 2006;367(9526):1903–12.

3. Hata J, Arima H, Zoungas S, et al. Effects of the endpoint adjudication process on the results of a randomised controlled trial: the ADVANCE trial. *PLoS One* 2013;8(2):e55807.

4. Ninomiya T, Donnan G, Anderson N, et al. Effects of the End Point Adjudication Process on the Results of the Perindopril Protection Against Recurrent Stroke Study (PROGRESS). *Stroke* 2009;40(6):2111–2115.

5. Pogue J, Walter SD, Yusuf S. Evaluating the benefit of event adjudication of cardiovascular outcomes in large simple RCTs. *Clin Trials* 2009;6(3):239–251.

6. Heckbert SR, Kooperberg C, Safford MM, et al. Comparison of self-report, hospital discharge codes, and adjudication of cardiovascular events in the women's health initiative. *Am J Epidemiol* 2004;160(12):1152–1158.

7. Calverley PMA, Anderson JA, Celli B, et al. Salmeterol and fluticasone propionate and survival in chronic obstructive pulmonary disease. *N Engl J Med* 2007;356(8):775–89.

8. Tashkin DP, Celli B, Senn S, et al. A 4-year trial of tiotropium in chronic obstructive pulmonary disease. *N Engl J Med* 2008;359(15):1543–54.

9. McGarvey LP, John M, Anderson JA, et al. Ascertainment of cause-specific mortality in COPD: operations of the TORCH Clinical Endpoint Committee. *Thorax* 2007;62(5):411–415.

10. McGarvey LP, Magder S, Burkhart D, et al. Cause-specific mortality adjudication in the UPLIFT® COPD trial: Findings and recommendations. *Respir Med* 2012;106(4):515–521.

11. Aaron SD, Fergusson D, Marks GB, et al. Counting, analysing and reporting exacerbations of COPD in randomised controlled trials. *Thorax* 2008;63(2):122–128.

12. Moy ML, Teylan M, Weston NA, et al. Daily Step Count Predicts Acute Exacerbations in a US Cohort with COPD. *PLoS One* 2013;8(4):e60400.

13. Donaldson GC, Seemungal TAR, Bhowmik A, et al. Relationship between exacerbation frequency and lung function decline in chronic obstructive pulmonary disease. *Thorax* 2002;57(10):847–852.

14. Miravitlles M, Ferrer M, Pont À, et al. Effect of exacerbations on quality of life in patients with chronic obstructive pulmonary disease: a 2 year follow up study. *Thorax* 2004;59(5):387–395.

15. Seemungal TAR, Donaldson GC, Paul EA, et al. Effect of exacerbation on quality of life in patients with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 1998;157(5):1418–1422.

16. Soler-Cataluña JJ, Martínez-García MÁ, Román Sánchez P, et al. Severe acute exacerbations and mortality in patients with chronic obstructive pulmonary disease. *Thorax* 2005;60(11):925–931.

17. Flegal KM, Brownie C, Haas JD. The effects of exposure misclassification on estimates of relative risk. *Am J Epidemiol* 1986;123(4):736–751.

18. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology - Third Edition. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.

19. Siebeling L, Riet G ter, Wal WM van der, et al. ICE COLD ERIC--International collaborative effort on chronic obstructive lung disease: exacerbation risk index cohorts--study protocol for an international COPD cohort study. *BMC Pulm Med* 2009;9:15.

20. Puhan MA, Siebeling L, Zoller M, et al. Simple functional performance tests and mortality in COPD. *Eur Respir J* 2013;42(4):956–63.

21. Frei A, Muggensturm P, Putcha N, et al. Five comorbidities reflected the health status in patients with chronic obstructive pulmonary disease: the newly developed COMCOLD index. *J Clin Epidemiol* 2014;67(8):904–11.

22. Siebeling L, Puhan MA, Muggensturm P, et al. Characteristics of Dutch and Swiss primary care COPD patients - baseline data of the ICE COLD ERIC study. *Clin Epidemiol* 2011;3:273–283.

23. Cazzola M, MacNee W, Martinez FJ, et al. Outcomes for COPD pharmacological trials: from lung function to biomarkers. *Eur Respir J* 2008;31(2):416–469.

24. Puhan MA, Behnke M, Devereaux PJ, et al. Measurement of agreement on health-related quality of life changes in response to respiratory rehabilitation by patients and physicians–a prospective study. *Respir Med* 2004;98(12):1195-202.

25. Schünemann HJ, Griffith L, Jaeschke R, et al. Evaluation of the minimal important difference for the feeling thermometer and the St. George's Respiratory Questionnaire in patients with chronic airflow obstruction. *J Clin Epidemiol* 2003;56(12):1170–6.

26. Guyatt GH, Berman LB, Townsend M, et al. A measure of quality of life for clinical trials in chronic lung disease. *Thorax* 1987;42(10):773–778.

27. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. *Acta Psychiatr Scand* 1983;67(6):361–70.

28. Wacholder S, Hartge P, Lubin JH, et al. Non-differential misclassification and bias towards the null: a clarification. *Occup Env Med* 1995;52(8):557–558.

29. Wang J, Nie B, Xiong W, et al. Effect of long-acting beta-agonists on the frequency of COPD exacerbations: a meta-analysis. *J Clin Pharm Ther* 2012;37(2):204–211.

30. Effing TW, Kerstjens HAM, Monninkhof EM, et al. Definitions of exacerbations: does it really matter in clinical trials on COPD? *Chest* 2009;136(3):918–923.

31. Trappenburg JCA, Deventer AC van, Troosters T, et al. The impact of using different symptom-based exacerbation algorithms in patients with COPD. *Eur Respir J* 2011;37(5):1260–1268.

32. Leidy NK, Wilcox TK, Jones PW, et al. Standardizing measurement of chronic obstructive pulmonary disease exacerbations. Reliability and validity of a patient-reported diary. *Am J Respir Crit Care Med* 2011;183(3):323–329.

Tables

Table 1: Total number of exacerbations per patient: patient 6-months recall and self-reports compared to adjudicated exacerbations by an adjudication committee (reference standard)

		Total number of centrally adjudicated exacerbations per patient													
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**Total**
Total number of self-reported exacerbations per patient	**0**	**127***	24	5	4	2	2	1	0	0	0	0	0	0	165
	1	26	**40**	5	2	1	3	0	0	0	0	0	0	0	77
	2	9	17	**10**	4	2	1	0	0	0	0	0	0	0	43
	3	3	6	7	**10**	2	3	2	1	0	0	0	0	0	34
	4	1	7	3	6	**2**	3	2	1	0	0	0	1	0	26
	5	0	3	5	4	0	**4**	1	1	0	0	0	0	0	18
	6	0	2	4	1	6	1	**2**	0	0	0	0	0	0	16
	7	0	2	2	0	2	0	0	**0**	0	0	0	0	0	6
	8	0	0	0	2	2	0	1	2	**1**	0	0	0	1	9
	9	2	0	0	1	0	0	0	1	1	**0**	0	0	0	5
	10	0	0	0	0	0	1	0	0	0	0	**0**	0	0	1
	11	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0
	12	0	0	0	0	0	2	0	0	2	0	0	0	**0**	4
	13	0	0	0	0	0	1	1	0	0	0	0	0	0	2
	14	0	0	0	0	1	0	0	0	0	0	1	0	0	2
	15	0	0	0	0	0	0	0	0	0	0	0	1	0	1
	Total	168	101	41	34	20	21	10	6	4	0	1	2	1	409

*Bold cells: number of correct self-reported exacerbations. E.g. see "Total number of exacerbations per patient, adjudicated in AC", column "1 exacerbation": 40 patients reported correctly that they had 1 exacerbation, 24 patients under-reported their exacerbations (they reported no exacerbation while they had had 1 according to the adjudication committee), 37 (77 minus 40) patients over-reported their exacerbations (reported they had had >1 exacerbations while they had had 1 according to the adjudication committee)

Table 2: Sensitivities and specificities of patient recall and self-reports and single experienced physician chart review compared to centrally adjudicated exacerbations by an adjudication committee (reference standard) (n=409)

Source*)	Number of patients	Sensitivity (%)†	Specificity (%)†
Patient recall and self-reports	409	84.2	75.6
Physician 1 CH	151	88.8	94.4
Physician 2 CH	151	93.8	97.2
Physician 3 CH	151	91.3	98.6
Physician 1 NL	258	94.4	96.9
Physician 2 NL	257	95.0	86.6
Physician 3 NL	258	96.3	94.8
Physician 4 NL	258	90.7	96.9

*CH = Switzerland; NL = The Netherlands
†Exacerbations categorised: no exacerbation vs. ≥1 exacerbation

Table 3: Multinomial regression model showing the relative risk ratios of the predictors for under- or over-reporting exacerbations compared to the base category correctly self-reporting exacerbations (n=403)

Variable	RRR*	Standard error	p-value	95% CI
Patients who under-reported exacerbations (n=74)				
Sex†				
Female	0.53	0.21	0.105	0.24 – 1.14
Age (years)	1.05	0.23	0.024	1.01 – 1.09
Education†				
Secondary school	0.98	0.45	0.973	0.40 – 2.43
Intermediate vocational	0.91	0.51	0.865	0.30 – 2.76
High vocational / university	1.25	0.83	0.737	0.34 – 4.62
Living situation†				
Living with partner	0.79	0.30	0.543	0.38 – 1.67
Living with children and/or partner	3.49	1.94	0.024	1.17 – 10.38
Working situation†				
Not working	1.00	0.47	0.998	0.40 – 2.53
Feeling thermometer‡	0.99	0.01	0.590	0.97 – 1.02
CRQ dyspnoea domain	1.01	0.14	0.934	0.77 – 1.33
CRQ fatigue domain	0.79	0.14	0.934	0.80 – 1.33
CRQ mastery domain	1.01	0.18	0.960	0.71 – 1.45
FEV_1 (litres)	1.00	0.38	1.000	0.48 – 2.10
HADS anxiety score	1.11	0.06	0.061	1.00 – 1.23
HADS depression score	0.89	0.06	0.103	0.78 – 1.02
Number of comorbidities	0.93	0.070	0.378	0.79 – 1.09
Number of exacerbations‡	2.16	0.23	<0.001	1.76 – 2.65
Patients who correctly reported exacerbations (base category, n=193)				
Patients who over-reported exacerbations (n=136)				
Sex†				
Female	0.74	0.22	0.317	0.42 – 1.33
Age (years)	1.01	0.02	0.617	0.98 – 1.04
Education†				
Secondary school	0.54	0.18	0.068	0.28 – 1.05
Intermediate vocational	0.61	0.26	0.244	0.27 – 1.40
High vocational / university	0.64	0.32	0.364	0.24 – 1.69
Living situation†				
Living with partner	0.78	0.22	0.379	0.45 – 1.36
Living with children and/or partner	1.50	1.68	0.364	0.62 – 3.63
Working situation†				
Not working	1.21	0.44	0.592	0.60 – 2.47
Feeling thermometer‡	0.99	0.01	0.290	0.97 – 1.01
CRQ dyspnoea domain	0.88	0.10	0.240	0.71 – 1.09
CRQ fatigue domain	0.97	0.13	0.807	0.74 – 1.27
CRQ mastery domain	1.12	0.16	0.429	0.85 – 1.47
FEV_1 (litres)	1.19	0.35	0.549	0.68 – 2.07
HADS anxiety score	1.07	0.04	0.083	0.99 – 1.17
HADS depression score	0.94	0.05	0.221	0.85 – 1.04
Number of comorbidities	1.07	0.06	0.271	0.95 – 1.20
Number of exacerbations‡	1.67	0.15	<0.001	1.39 – 2.00

Abbreviations and scores: FEV_1: Forced expiratory volume in 1 second; CRQ=Chronic Respiratory Questionnaire (scores: 0-7/maximal to no impairment); HADS: Hospital Anxiety and Depression Scale (scores: 0-21/no to most severe anxiety/depression); feeling thermometer (score: 0-100/worst to best health status). *RRR = Relative risk ratio. RRR describes the multiplicative effect of a unit increase in each predictor on the odds of over- or under-reporting instead of correctly self-reporting exacerbations (base category). †Comparison against following categories: Sex: male / Education: lowest level / Living situation: living alone / Working

situation: still working. ‡In the bootstrap stability analysis (500 replicates, selection p<0.1, threshold of bootstrap inclusion fraction (BIF) 67%) the inclusion fractions for the variables *number of adjudicated exacerbations during study* (BIF=100%) and *feeling thermometer* (BIF=67.7%) exceeded the 67% threshold.

Figure legends:

Figure 1: Recalculation of meta-analysis of long-acting bronchodilators vs. placebo trials with exacerbations as the outcome, based on four scenarios of misclassification and derivation of sample size requirements with and without use of adjudication committee

Legend: The figure shows four examples of sample size calculations that compare calculations using the originally published pooled OR (0.81) of the meta-analysis of LABA vs. placebo trials (where self-reports of exacerbations were used) against ORs attained from recalculated meta-analyses that were adjusted for four assumed scenarios of misclassifications (detected when adjudication committees were used), for a low risk (10% risk over 1 year) and a high risk (50% risk over 1 year) population situations: Example a) sensitivity and specificity of 80% and 70%, respectively, results in OR 0.59; b) 84%/76% in OR 0.65 (such as detected in our study), c) 80%/85% in OR 0.69, d) 95%/85% in OR 0.73.